Mining Knowledge Components From Many Untagged Questions

Neil L Zimmerman
McGraw-Hill Education
281 Summer St
Boston, MA 02210
neil.zimmerman@mheducation.com

Ryan S Baker
Graduate School of Education
University of Pennsylvania
3700 Walnut Street
Philadelphia, PA 19104
rybaker@upenn.edu

ABSTRACT
An ongoing study is being run to ensure that the McGraw-Hill Education LearnSmart platform teaches students as efficiently as possible. The first step in doing so is to identify what Knowledge Components (KCs) exist in the content; while the content is tagged by experts, these tags need to be re-calibrated periodically.

LearnSmart courses are organized into chapters corresponding to those found in a textbook; each chapter can have anywhere from about a hundred to a few thousand questions. The KC extraction algorithms proposed by Barnes [1] and Desmarais et al [3] are applied on a chapter-by-chapter basis. To assess the ability of each mined matrix to describe the observed learning, the PFA model of Pavlik et al [4] is fitted to it and a cross-validated AUC is calculated. The models are assessed based on whether PFA’s predictions of student correctness are accurate.

Early results show that both algorithms do a reasonable job of describing student progress, but matrices with very different numbers of KCs fit observed data similarly well. Consequently, further consideration is required before automated extraction is practical in this context.

CCS Concepts
• Information systems → Data mining; • Applied computing → Computer-assisted instruction; Learning management systems; E-learning;

Keywords
Knowledge Components; Knowledge Tracing; Data Mining

1. INTRODUCTION
The LearnSmart platform1 is an adaptive learning system that follows along with the textbook used in a course; students are given questions corresponding to topics in each chapter in an order prescribed by the system, and must correctly answer a certain number of questions from each topic before the platform has deemed that they have mastered the chapter. LearnSmart is very widely used; in an average month, there are more than 100,000,000 distinct student-question interactions.

The actual learner experience in LearnSmart is dependent on metadata produced by subject matter experts, namely: what topics exist in each chapter, and how much evidence does a correct response to a given question contribute to proving that a student has mastered the topic? This paper discusses part of a broad effort to confirm the validity of this metadata, first by mining knowledge components from student interactions and then (as part of a future work), to trace the acquisition of these components and confirm that students are not provided with too little or too much practice on each topic.

This paper focuses on a single chapter from a single course teaching the Spanish language to English speakers. In this chapter there are 133 questions, given to roughly 11,000 students in a sequence deemed appropriate by the adaptive platform. In total, there are approximately 750,000 distinct student-question interactions in this dataset.

2. METHODOLOGY
In this work, we attempt to extract KCs from “chapters” automatically, as the topic and skill tagging is absent in some textbooks, and the sheer volume of texts makes manual tagging impractical. This approach assumes that KCs will not be found across chapters; while this assumption is unlikely to be formally true, the pedagogy of this platform is such that we want to measure KC progress on a chapter-by-chapter basis.

2.1 Imputing missing features
While students do not have any control over the sequence of questions that they seem the adaptive platform generally deems students to be finished with a given chapter after showing them roughly half the questions. While the algorithm of Barnes [1] can be easily adapted to work with missing data, regression techniques like non-negative matrix factorization (NMF) cannot.

Therefore, for each question, a logistic model is built using all students who did attempt the question as training data, where the features are the students’ interactions on all other questions in the chapter, using one-hot encoding to turn

1http://www.mheducation.com/highered/platforms/learnsmart.html
3. CONCLUSIONS AND NEXT STEPS

There are two immediate and parallel paths that need to be pursued: ensuring that the flat AUC is not an artifact of either the NMF for factoring or PFA for scoring (e.g., by comparing with the algorithm in [1] or with Bayesian Knowledge Tracing [2], or by removing the difficulty component in PFA).

If this flat AUC is a robust feature of the data, then NMF should be revisited using matrix factorization technique that does not require imputing missing data, at least to confirm that it produces q matrices similar to those produced when imputing data. Additionally, comparing the 2-means method of turning non-binary matrices into binary matrices, used by Desmarais, with explicitly binary matrix factorization techniques (e.g. [7]), would help ensure that using NMF to produce binary matrices produces comparable results.

4. ACKNOWLEDGMENTS

The authors would like to thank the MHE LearnSmart team for providing access to data used in this analysis.

5. REFERENCES